XAI IN IMDB MOVIE REVIEW SENTIMENT ANALYSIS

Challenging Assignments and Mini Projects (CHAMP)

submitted as part of the course
Explainable Artificial Intelligence
BCSE418L
School Of Computer Science and Engineering
VIT Chennai

FALL 2023-2024

Course Faculty : Dr. B Radhika Selvamani
Submitted By
Nikhil Phaneendra (21BAI1345)
Kalyan Chakravarthy Y (21BAI1759)
Aasrith Dogiparthi (21BAI1702)
Shobith Paripalli (21BAI1722)

Abstract

In natural language processing (NLP), sentiment analysis serves as a crucial tool for extracting valuable insights from textual data, particularly in large-scale datasets like IMDB movie reviews. It enables businesses, researchers, and other stakeholders to gauge public sentiment and make informed decisions based on the extracted patterns. However, the application of sentiment analysis is not without challenges. Many of the predictive models employed, particularly those based on deep learning, suffer from a lack of interpretability. These models often function as "black boxes," delivering accurate predictions but offering little insight into the reasoning behind their outputs. This lack of transparency can lead to mistrust and hesitancy in adopting these systems for decision-making.

To address this issue, our study integrates Explainable Artificial Intelligence (XAI) techniques to make sentiment analysis more transparent and interpretable. Specifically, we utilize SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to shed light on the decision-making processes of deep learning models. These techniques provide local and global explanations of model predictions, enabling users to understand which aspects of the data influence the sentiment classification and how these contributions affect the overall outcome.

Our approach involves applying these XAI techniques to advanced deep learning frameworks, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. These models are well-suited for sentiment analysis due to their ability to capture complex relationships in text, such as semantic dependencies and sequential context. The primary objective of this research is to balance high predictive accuracy with enhanced interpretability, ensuring that the models remain reliable while also offering insights into their inner workings.

Preliminary findings reveal that incorporating XAI techniques leads to significant improvements in user trust and comprehension of the model's outputs, without compromising performance. For instance, users can now identify specific words or phrases that strongly influence sentiment predictions, such as terms with high positive or negative connotations. This level of interpretability not only increases confidence in the results but also helps stakeholders validate the predictions against domain-specific knowledge or expectations.

This study highlights the transformative potential of XAI in sentiment analysis and beyond. By demystifying the decision-making process of AI models, we pave the way for more transparent, accountable, and ethical applications of machine learning in fields like media analysis, customer feedback evaluation, and social media monitoring. The ability to make AI systems both accurate and interpretable represents a significant step forward in bridging the gap between technical advancements and their practical adoption, empowering stakeholders with tools for more informed, data-driven decision-making.

Contents

1 Introduction

Relevance of Sentiment Analysis Sentiment analysis:
Sentiment analysis plays a key role in natural language processing (NLP). It enables us to get great insights into how people feel, all through their words. Researchers and businesses can figure out what the public thinks about lots of different topics. This includes everything from product reviews to movie critiques. The IMDB dataset, filled with movie reviews, is a fantastic resource. It helps train AI models to tell if sentiments are positive or negative. This whole process is extremely important for understanding how consumers react and for shaping marketing strategies, customizing content, and making smarter business decisions.

Increasing Demand for Explainability in AI:
As AI gets fancier and finds its way into many areas, the need for transparency is growing. The old black-box way - where users can't see what's going on inside AI models, causes real problems. In many fields, knowing why a decision was made is key. If things are unclear, it can affect how much users trust or accept these technologies. That makes it hard to fully use all the great advantages AI can offer.

Objectives of the Project:

This project, "Explaining Emotions," aims to tackle challenges by mixing Explainable AI (XAI) with sentiment analysis of the IMDB dataset. The goals here are:

Enhanced Transparency: Using XAI methods like SHAP and LIME, the project plans to peel back layers of AI decision-making. We want to make these decisions clear and understandable for users.

Maintain Performance: It's important that adding explainability doesn't mess up how accurately sentiments are classified. This study will check how XAI affects model performance to make sure we don't lose effectiveness while gaining transparency.

Foster Trust: Giving clear, simple explanations on how sentiments get classified can help build trust among users. They should feel confident in validating and relying on the decisions that AI makes.

Broaden Accessibility: It's essential to make AI and its choices easy to grasp for those who aren't experts.

This study aims at democratizing AI tech so it's accessible and usable by more people without needing technical expertise.

2 Related Works

The paper provides a comprehensive survey of the current state of sentiment analysis, emphasizing the dominance of deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) in achieving state-of-the-art performance. However, it highlights a critical challenge associated with these models: their "black-box" nature, which leads to significant issues in interpretability and the ability to understand and justify their decision-making processes. To address these challenges, the paper explores the field of eXplainable Artificial Intelligence (XAI), which aims to make AI models more interpretable. The survey discusses various approaches and definitions of explainability, particularly in the context of sentiment analysis, and examines the ongoing debate between model performance and interpretability. Additionally, the paper identifies the need for further research into explainable models for multilingual sentiment analysis, suggesting that future efforts could focus on leveraging Graph Neural Networks to develop dependency-based rules for different languages.

This paper has significant advances in the field of explainable artificial intelligence (XAI) within supervised learning. It highlights key works that have explored the ethical and trust-related challenges posed by the opacity of deep learning models, particularly in sensitive applications like healthcare and finance. The survey discusses the emergence of various interpretability methods aimed at making AI systems more transparent and accountable. It also examines previous surveys and reviews, emphasizing the growing importance of XAI in ensuring the reliability and fairness of machine learning models. Additionally, the paper addresses the broader implications of these developments for artificial general intelligence (AGI) and ethical AI.

The literature survey in this paper reviews various approaches to automatic sarcasm detection, including rule-based, statistical, machine learning, and deep learning techniques. It highlights the challenges of detecting sarcasm, especially in single sentences, and the importance of contextual information in improving accuracy. The survey discusses the evolution of sarcasm detection methods from focusing on isolated utterances to incorporating broader contextual cues like social-graph, temporal context, and user profiles. It also examines recent advancements in multimodal sarcasm detection, integrating text with images and other media. The need for more sophisticated methods to capture contextual incongruity in dialogues and conversational threads is emphasized, pointing to ongoing challenges and future research directions.

The literature survey in this paper explores the rapidly evolving field of Explainable Artificial Intelligence (XAI), addressing the critical need for transparency in AI systems. It reviews key contributions and approaches to XAI, including efforts by academia, industry, and government organizations like DARPA, highlighting the importance of fairness, accountability, and transparency in al-

gorithmic decision-making. The survey discusses the growing body of research focused on creating explainable models without compromising performance, emphasizing the interdisciplinary nature of XAI. It also identifies gaps in formalism and the limited study of the human role in existing approaches, suggesting areas for future research and development.

The literature survey in this paper delves into the recent advancements in Aspect-based Sentiment Analysis (AbSA), highlighting key challenges and innovations in the field. It reviews methods for Aspect Extraction, emphasizing both explicit and implicit aspects, and discusses various approaches for Aspect Sentiment Analysis that focus on improving sentiment classification accuracy through interactions and contextual semantic relationships between data objects. The survey also addresses the dynamic nature of Sentiment Evolution over time, considering social characteristics and self-experience as influential factors. Additionally, the paper categorizes recent research solutions based on their contributions to different phases of AbSA, providing a comprehensive overview of the state-of-the-art in this domain.

3 Design

The following outlines how data is processed, the model is trained and tested, and how XAI is used to interpret the model.

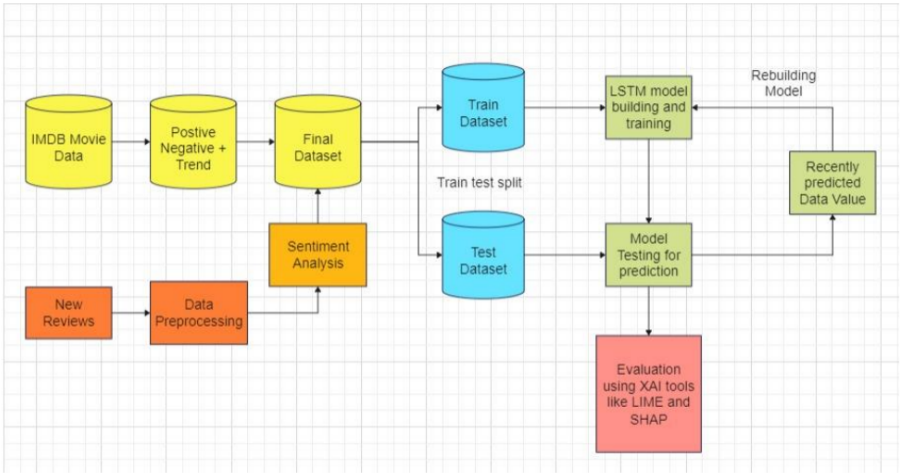

Figure 1: Architecture

The above figure illustrates a sentiment analysis pipeline for movie reviews from IMDB. It begins by collecting IMDB movie data, which is then processed to classify reviews into positive, negative, and trend categories. This results in

a final dataset that undergoes a train-test split for model development. A Long Short-Term Memory (LSTM) model is built and trained on the training set, followed by testing on the test set for sentiment prediction. The model's performance is evaluated using explainability tools like LIME and SHAP, allowing for deeper insights into its predictions. If needed, the model is rebuilt based on newly predicted data to improve accuracy. Additionally, the system integrates new reviews, processes them, and updates the sentiment analysis model, enabling continuous learning and improvement.

4 Procedure

4.1 Data Collection and Pre-processing

The process begins with obtaining the IMDB movie reviews dataset, which contains user-written reviews categorized by sentiment (positive, negative, or neutral). The dataset can be sourced from platforms like Kaggle or gathered by directly scraping reviews from the IMDB website. Once the dataset is acquired, the data undergoes a cleaning process. This involves removing unnecessary characters, such as punctuation, HTML tags, and special symbols, to ensure clean text input. Additionally, the text is converted to lowercase, and stop words (e.g., "the," "is") are eliminated to focus on meaningful words that contribute to sentiment analysis. If the dataset lacks predefined sentiment labels, tools like VADER or TextBlob can be employed to classify reviews into positive, neutral, or negative sentiment categories.

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

Figure 2: Data

Figure 3: Word Cloud

4.2 Building the Sentiment Analysis Model (Bidirectional LSTM)

To build a sentiment analysis model, the dataset is first prepared for training by splitting it into training and test sets, typically using an 80/20 or 70/30 ratio. The text reviews are then converted into numerical representations through techniques such as word embeddings and tokenization, which are essential for processing sequential data. The core of the model is a Bidirectional LSTM (BiLSTM), designed to capture both forward and backward dependencies in the text. Its architecture includes an embedding layer to map words to vector embeddings, a BiLSTM layer to process input sequences bidirectionally, dense layers for classification, and a softmax activation function for multi-class output. Dropout layers are incorporated to regularize the model and prevent overfitting. The BiLSTM model is trained using the training dataset, leveraging loss functions like categorical cross-entropy and optimizers such as Adam. Once trained, the model's performance is evaluated on the test set using metrics like accuracy, precision, recall, and F1-score. The trained model is then saved for sentiment prediction and further use in explainability studies.

```
model.summary()
```

```
Model: "sequential"
```

Layer (type)	Output Shape	Param #
embedding (Embedding)	?	0 (unbuilt)
bidirectional (Bidirectional)	?	0 (unbuilt)
global_max_pooling1d (GlobalMaxPooling1D)	?	0 (unbuilt)
dense (Dense)	?	0 (unbuilt)
dropout (Dropout)	?	0 (unbuilt)
dense_1 (Dense)	?	0 (unbuilt)

Figure 4: Model Architecture

4.3 Integrating Explainable AI (XAI) Tools

Explainability is introduced using tools like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations). LIME explains individual sentiment predictions by perturbing input data and observing the BiLSTM model's responses. This approach highlights specific words or phrases that influenced the model's predictions. Similarly, SHAP provides both

global and local explanations. It assigns importance values to features (words) based on their contributions to the model's predictions. SHAP visualizations enable an understanding of the model's overall behavior (global explanations) and the reasoning behind specific predictions (local explanations). Together, these tools enhance the interpretability of the BiLSTM model.

4.4 Evaluating Model and Explainability

The evaluation phase involves analyzing the explanations provided by LIME and SHAP for individual reviews. Visual tools are used to identify and highlight key words or phrases that significantly influenced the sentiment predictions. The model's performance metrics—accuracy, precision, recall, and F1-score—are assessed alongside the explanations to ensure they align with human intuition. A comparison is made between the model's predictive accuracy and its explainability to determine whether the insights provided by XAI tools are consistent with the review content. By combining performance metrics with visual explanations, the transparency and interpretability of the BiLSTM model are thoroughly evaluated, ensuring its reliability and utility for sentiment analysis.

5 Results and Discussion

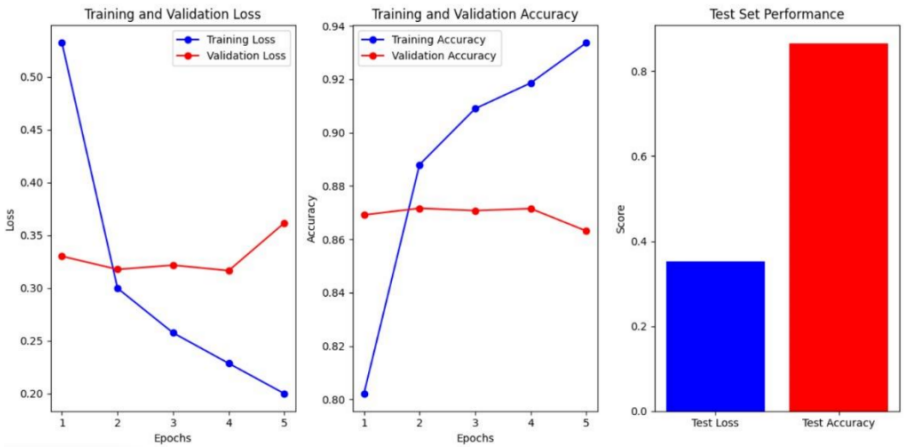

Figure 5: Model Results

The results depicted in the above charts indicate that the model is learning effectively on the training data, as shown by the significant decrease in training loss and the consistent increase in training accuracy, reaching almost 94% by the final epoch. The validation loss remains relatively flat and starts to increase after the third epoch, while the validation accuracy plateaus around 87–88%,

showing little improvement. The test set results, with a lower test loss and high test accuracy above 85%, indicate that the model performs well on the unseen test data.

As for the model, as shown in the figure below, we obtain explanation features. The word *terrible* is mostly related in a negative context and hence is associated with a negative sentiment. Similarly, *best* is commonly found in sentences that depict positive sentiment. From the table below, we can interpret our data and understand the biases in the dataset. Our model is no longer a black box, as the customers can gain insight into the system through a single graph. Thus, we achieve explanations along with insights from the data.

5.1 LIME Results

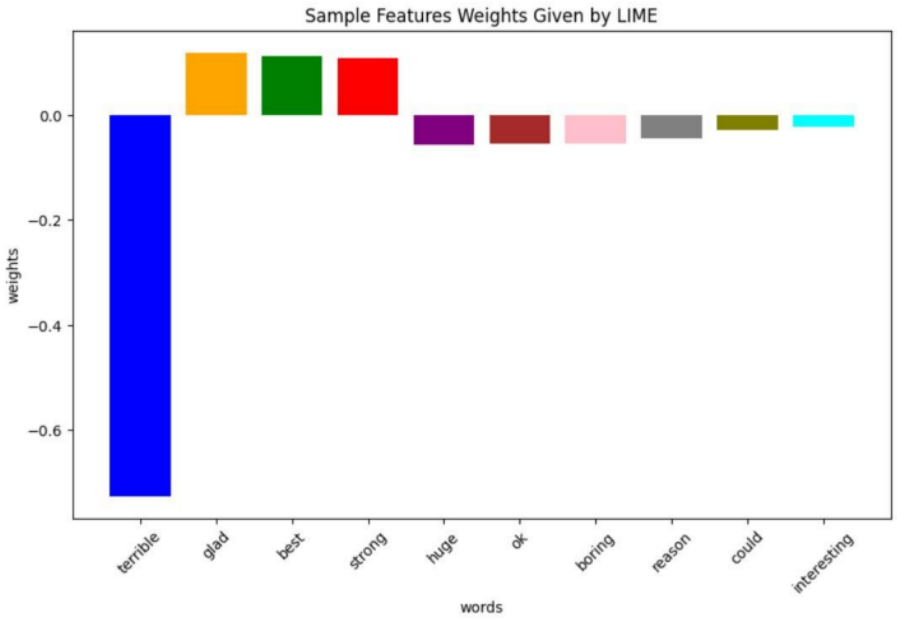

Figure 6: Results of LIME Tool

	Test Words	Weights or biases (approx)
0	terrible	-0.726358
1	glad	0.118465
2	best	0.111906
3	strong	0.107521
4	huge	-0.056418
5	ok	-0.055174
6	boring	-0.055108
7	reason	-0.043741
8	could	-0.027866
9	interesting	-0.021979

Figure: Results of LIME Tool

Figure 7: LIME

The image above depicts the analysis of a text using a sentiment classification model. The prediction probabilities indicate an 86% likelihood that the text is negative and a 14% chance that it is positive. Key words contributing to the negative sentiment, such as *terrible* (with the highest negative contribution), *boring*, and *reason*, are highlighted. Positive words like *glad*, *best*, and *strong* contribute to a much smaller portion of the sentiment analysis. In the highlighted text, the negative tone is reinforced by the repeated use of the word *terrible* to describe events, matches, and the overall experience, along with terms like *boring* and *fat man*. Positive words like *glad* and *interesting* seem to have little impact on altering the overall negative sentiment. The text reflects a critical perspective of certain aspects of the event, emphasizing disappointment regarding specific matches and performers.

Figure 8: SHAP Results

The image above shows a SHAP (SHapley Additive exPlanations) plot illustrating the contribution of specific words to the model's sentiment prediction. The base value starts at 0.5277, and the final prediction score, $f(x)$, is 0.14,

indicating a predominantly negative sentiment. Words that pushed the prediction towards a more positive sentiment are displayed on the left in red, while words that contributed to a more negative sentiment are on the right in blue.

Words like *matches*, *marty*, *deserves*, and *best* have higher SHAP values in red, suggesting they influenced the model towards a more positive direction. For example, *matches* contributed 4.011, and *marty* 4.723 to raising the sentiment score. However, words such as *terrible*, *boring*, *huge*, and *ok*, shown in blue, pushed the sentiment towards negativity. The repeated use of *terrible* contributed significantly to lowering the sentiment, with each occurrence of *terrible* reducing the score by 0.255.

Overall, despite some positive words, the negative words, especially *terrible*, had a stronger impact, leading to the overall negative prediction of the text.

6 Conclusion

Sentiment analysis of IMDb reviews significantly impacts how viewers perceive movies and shows, as it reflects collective opinions about content quality. Predicting movie sentiments based on reviews can help viewers and creators gauge audience reception and improve recommendations. Our proposed model not only provides accurate sentiment predictions but also explains the reasoning behind these predictions using Explainable AI (XAI) through SHAP analysis, making the output more transparent and interpretable. Future research could explore automated sentiment predictions from real-time review platforms and extend to multilingual review analysis. Additionally, integrating visual sentiment indicators could enhance user engagement and understanding. This model could also serve as a recommendation tool for streaming services to personalize content for users.

References

[1] A. Diwali, K. Saeedi, K. Dashtipour, M. Gogate, E. Cambria and A. Hussain, "Sentiment Analysis Meets Explainable Artificial Intelligence: A Survey on Explainable Sentiment Analysis," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 837-846, July-Sept. 2024, doi: 10.1109/TAFFC.2023.3296373.
Keywords: Sentiment analysis; Analytical models; Computational modeling; Task analysis; Deep learning; Predictive models; Artificial neural networks; Explainability; Interpretability.

[2] F. K. Došilović, M. Brčić and N. Hlupić, "Explainable Artificial Intelligence: A Survey," *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2018, pp. 0210-0215, doi: 10.23919/MIPRO.2018.8400040.
Keywords: Predictive models; Machine learning; Support vector machines;

Decision trees; Supervised learning; Optimization; Explainable artificial intelligence; Interpretability; Explainability; Comprehensibility.

[3] A. Kumar, S. Dikshit, and V. H. C. Albuquerque, "Explainable Artificial Intelligence for Sarcasm Detection in Dialogues," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 2939334, 2021.
Keywords: Explainable artificial intelligence; Sarcasm detection; Machine learning; Interpretability.

[4] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138-52160, 2018, doi: 10.1109/ACCESS.2018.2870052.
Keywords: Conferences; Machine learning; Market research; Prediction algorithms; Machine learning algorithms; Biological system modeling; Explainable artificial intelligence; Interpretable machine learning; Black-box models.

[5] A. Nazir, Y. Rao, L. Wu and L. Sun, "Issues and Challenges of Aspect-Based Sentiment Analysis: A Comprehensive Survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 845-863, April-June 2022, doi: 10.1109/TAFFC.2020.2970399.
Keywords: Sentiment analysis; Social networking (online); Data mining; Machine learning; Task analysis; Tools; Aspect; Computational linguistics; Deep learning; Sentiment evolution; Social media.